A Potential Quality Assurance of the Re-coding to NACE Rev. 2.1, Combining LLMs and Manual Coding

Jacob Kasche[1], Wictoria Widén[1], Kira Gylling[1], Gustaf Strandell[1]

[1]Statistics Sweden, Sweden

Abstract

Implementing NACE Revision 2.1 is demanding for many European countries. A major part of the transition is the re-coding of units in the Business registers. Previously, the re-coding process has mostly been done using surveys and manual coding, which often result in large costs. Quality demands on NACE are high; hence quality needs to be high in the re-coding process. In previous quality assurance processes, several coders repeated the re-coding i.e., reconciliation. Because of budget restrictions, it may not be feasible to perform this process on the entire nomenclature.

Because of the increased performance of large language models (LLMs), several countries investigate the possibilities of using LLMs to decrease manual coding. However, the model approach does not only increase the possibilities to lower the use of manual resources. It also facilitates the development of effective quality assurance.

In this paper, a potential quality assurance process, which focuses on combining manual labour with models e.g., LLMs, is presented. The quality assurance process includes: 1. Model inference; 2. Design inference with auxiliary information; 3. Manual coding supported by models; 4. Re-use of manually coded data. Methodologies necessary for each step are presented and the workflow is illustrated with examples from Statistics Sweden. Lastly, the paper discusses the quality assurance process and how it may facilitate an effective transition in the current and upcoming revisions of NACE for an NSI.

Keywords: NACE, LLM, Quality assurance, Coding

1. Introduction

NACE Revision 2.0 is the current statistical classification of economic activities in the European Union which assures a common standard for European statistics. Implementing NACE Revision 2.1 is demanding for European countries and this paper discusses the challenges in the context of Statistics Sweden. A major part of the transition is the re-coding of units in the Business register.

In previous quality assurance processes at Statistics Sweden, several human coders repeated the re-coding. Because of budget restrictions, it may not be feasible to perform this process over the entire nomenclature for the current revision. This was not possible at the former revision to NACE 2.0 either. Only certain codes and units, which were considered difficult to code manually, were controlled. In general, manual coding is considered to be of stable quality.

Because of the increased performance of large language models (LLMs), several countries investigate the possibilities of using LLMs to decrease manual coding. The model approach does not only facilitate lower use of manual resources. It may also be used to develop effective quality assurance since a model may provide uncertainty indicators for predicted values for the whole population. In general, process data from a model may be used to measure and improve the result in the coding process, although one must take care as the use of models may introduce additional uncertainty.

An increasingly automated coding process may rise questions regarding the quality of the outcome. Quality demands on NACE are high in Sweden because of the diverse use of the Business register for statistical and administrative purposes. Wallgren and Wallgren (2022) presents four quality concepts of register surveys: 1. Input data quality, 2. Production process quality, 3. Output data quality, 4. Quality of statistical estimates.The diverse purposes rise demands regarding microdata i.e., Concept 3, and statistical estimates i.e., Concept 4. Hence, the quality needs to be high regarding distributions as well as at the unit level. The first two concepts are also relevant since the errors are inherent in the succeeding quality concepts.

The purpose of this paper is to present and discuss a quality assurance process when applying a combined coding process through manual and model coding. The process is presented for use in the current NACE revision. First, we present and discuss data and coding methods. Secondly, we describe the quality assurance process, including methods, prioritizations, and integration with the coding methods. Lastly, the quality assurance process is discussed.

2. Data and Coding methods

2.1 Data

The former revision to NACE Rev. 2.0 set out from the NACE Rev. 1.1 code. The main data source used for re-coding were textual descriptions of the units' economic activities, which were compared with the explanatory notes and examples of activities of the NACE Rev. 2.0 codes. The process also utilized accounting, geographical, and occupational data and relied heavily on subject matter knowledge of NACE and data. The above-mentioned types of data are suggested as appropriate for use in the ongoing revision.

In the handbook of implementation of NACE Rev. 2.1 (Eurostat, 2023) they underline the importance of quality indicators for data. We suggest using the timestamp of when the NACE Rev. 2.0 code was collected, the timestamp of the textual data, and the difference in time between the NACE Rev. 2.0. code and textual data, as quality indicators for the validity aspect; see also Subsection 3.1.1. Quality of training data can also be regarded through concepts used in the total survey error framework, mainly measurement error, coverage error and sampling error. These types of errors may be reduced during the learning phase, depending on the design of the model, or estimated during the prediction phase (UNECE Machine Learning Group, 2022).

2.2 Coding methods

In previous transitions to a new NACE nomenclature, the re-coding process consisted of three methods: surveys, internal manual coding, and automatic coding. The latter encompassed those units for which the previous and revised NACE codes had a one-to-one relationship and units which were not possible to code manually; the latter are further described in Section 2.2.4. However, to train an automatic coding model is not straightforward, due to the missing values for the response variable, i.e., the NACE Rev. 2.1 codes. A standard approach is to manually code a sample of data and then train a supervised model on this sample. This however requires additional resources and may not be efficient. We suggest using a combination of several methods to facilitate an effective coding process.

2.2.1 Manual coding

Manual coding may be performed internally at Statistics Sweden or externally by a business representee. Furthermore, internal coding may consist of coding done by one or several coding staff or an expert i.e., a person with subject matter knowledge. In previous revisions, external coding was done through surveys or other direct contacts with the businesses. Statistics Sweden will in the current transition use an online platform for collecting codes directly from the business representees. The platform is a well-established platform when businesses are

in contact with Swedish governments, e.g., at registration. Nevertheless, the use of the platform is voluntarily and hence the total respondent rate is assumed to be low.

2.2.2 Large Language Models

Since the previous revision of NACE, the availability of computational power and Large Language Models (LLMs) has increased, and several statistical institutes are investigating their usefulness in the coding process. However, using automatic solutions in combination with textual data for industrial coding is not new. In 1988, Statistics Canada presented a string-matching system which compares definitions of industrial codes with the business descriptions (Eurostat, 2023). Overall, the approach to compare definitions with descriptions is the same when using LLMs in the coding process. The difference lies in how the systems compare text. Statistics Canada's system compares similarity in letters and words, while LLMs compare semantic similarity. Hence, the definitions do not need to be as comprehensive and the descriptions not as formal as in the old system.

The utility of LLMs may vary from mainly calculating similarity between words e.g., BERT (Devlin, Chang, Lee, & Toutanova, 2018), to generating answers e.g., GPT-4 (OpenAI,2024). For our specific task, the former is sufficient. To increase the utility and performance in the LLM case, Statistics Sweden uses a version of BERT finetuned to Swedish, KB-BERT (Malmsten, Börjeson, & Haffenden, 2020). To measure the similarity between a business description and the definition of a NACE code we use the cosine similarity, i.e., the angle between the corresponding word vectors.

2.2.3 Clustering

Clustering techniques may also be useful. For example, they may be used on textual data to find units which have not received a new NACE Rev. 2.1 code but are similar to units which already received a specific code or to find units which stand out.

2.2.4 Imputation

For units with missing textual data or textual data of low quality, a different model is necessary. Primarily, this model uses other variables from, e.g., geographical, occupational, and accounting data. This process is closer to an imputation process since it relies on the likelihood of a NACE-code given the explanatory variables instead of comparing textual data with the code definitions. However, imputations may be performed in several ways. In the previous transition, remaining units which were not possible to code and for which the previous and new codes had a one-to-many relationship were imputed by the most common choice. Mainly, this approach may satisfy quality demands regarding microdata but neglect the distributions. Therefore, we suggest training a supervised classifying model e.g., a simple Neural Network

or a Random Forest model. Training data for the model originates from the methods described in Subsections 2.2.1-2.2.3. The suggested quality indicator from this model is an uncertainty value, e.g., the ratio of agreeing predictions in a Random Forest.

3. Quality Assurance Process

Quality demands regarding NACE-codes are broad due to the diverse use of the variable both for administrative and statistical purposes. The demands may differ, both regarding units and codes. In the latest transition to NACE Rev. 2.0, the re-coding process consisted mainly of manual coding and direct contact with the businesses. These alternatives are costly both with respect to response burden and monetary cost. Codes obtained by contacts with enterprises were regarded as the ground truth; however, the manual coding process was assumed to add uncertainty. An attempt to measure the uncertainty was made for a few selected NACE-codes and a relatively small sample. Nevertheless, a manual coding process can be assumed to have a stable and high quality.

In the current transition, which features an increase of coding performed by models, it is of essence to have an extensive quality assurance process, which is flexible enough to serve different quality demands. We therefore present different methods, which may be applied to different cases utilizing a priority system. This facilitates an effective allocation of manual resources. Quality assurance in the coding process through continuous improvements is not new, for example, a workflow is presented in Biemer and Lyberg (2003). Instead, the contribution of the present paper is to show how this may be done more effectively using process data from the coding process and the input data i.e., the quality indicators.

3.1 Definitions

3.1.1 Quality indicators and Quality measures

We define quality indicators as variables which indicate the quality on a NACE-code belonging to a given unit. Indicators may relate to input data quality or be the result from a model. Regarding input data quality, the paper focuses on the timeliness of data; see Subsection 2.1. Regarding model results, the paper focuses on the uncertainty of the prediction from the models which were described in Subsection 2.2.

In addition, we define quality measures as statistics related to a specific group of objects. Related to the diverse use of the Swedish business register as both an administrative and a statistical register, the quality measures need to describe both the error in microdata and possible effects on statistical estimates. Our suggestion is to cover this by measuring accuracy and absolute percentage error (Wallgren & Wallgren, 2022). The absolute percentage error is suggested for the following statistics: the total number of units, the total number of employees,

and total revenue. These measures cannot be calculated without manually coding every unit several times, but it is possible to estimate quality indicators using the methods described in Subsection 3.2.

3.1.2 Coding error

To calculate or estimate the quality indicators, definitions of erroneous values and true values are necessary. At Statistics Sweden, these are defined through double independent coding, followed by a reconciliation if the coders disagree. In addition, we define codes collected from business representees or coding by internal experts as true codes.

3.2 Methods

3.2.1 Model inference

Method 1, i.e., model inference, estimates the quality measures by using a model to estimate the probability of a unit, Y_i, belonging to a specific NACE-code, a, given the quality indicators i.e.,$\Pr(Y_i = a|Quality\ indicators)$. For predictive models this probability is often estimated when testing the model after the training phase on new unseen data i.e., test data. Moreover, it is recommended to resample out-of-sample data to achieve an estimated variance. Quality indicators may be used to improve or obtain a more precise quality measure. For example, France's NACE model uses a quality indicator as a threshold to increase the estimated accuracy (Faria & Seimandi, 2023). The estimate of the quality measure is only valid for new data under certain properties, e.g., that test data are representative for new data. Consequently, given the risk that these properties do not hold or the fact that estimates can have wide confidence intervals, it may not be feasible to use Method 1 for the most important cases.

3.2.2 Design inference

Method 2, i.e., design inference, estimates the quality measures from a sample coded by manual coding and hence the inference may be referred to as design based. For an effective use of manual resources when measures are estimated from a sample it is recommended to use the quality indicators as auxiliary information. The relationship between the quality measures and the quality indicators cannot be assumed to be linear and hence its necessary to use a non-linear model. Sande and Zhang (2021) show how to use a non-linear model as a GREG estimator and how to correct the bias of the estimator. The use of quality indicators is not limited to the estimation phase. They may be used in the sampling process as stratifying variables or as size variables in a probability proportion to size (PPS) sample. This would facilitate sampling the most uncertain objects to a greater extent.

3.2.3 Manual coding supported by models

Method 3 uses models to support an effective manual coding. This may be done through selecting the most uncertain predictions according to the quality indicators, or highly influential units according to the estimated quality measure, e.g., legal units with a high amount of revenue. Method 3 may also include unsupervised methods e.g., clustering techniques, to handle several similar objects at the same time and hence receive a more consistent coding. Method 3 may also be used to select specific outliers, which may be referred to as diversity sampling (Mosqueira-Rey et al., 2023). Primarily, the manual coding should be done by experts or by reconciliation.

3.2.4 Re-use of manually coded data

With Method 4, manually coded data are reused to evaluate and improve the coding process. Manually coded data may come from the above-mentioned methods or other internal and external coding. Coded units may come both from probability and non-probability samples. The latter are not as suitable for estimating the quality measures as Method 2. Instead, it suggests focussing on finding systematic errors in the coding process or improving the models with additional data. Hence, this again emphasises sampling data which has high uncertainty for manual coding. However, an extensively updated model may induce new uncertainty due to overfitting. It is recommended to update only obvious systematic errors or simply re-running the existing pipeline for model training utilizing the new data.

To summarize, the methods in the quality assurance process focuses both on how to estimate the quality measures with different costs of manual coding i.e., methods 1-2, and how to improve the quality of coding i.e., methods 3-4

3.3 Priority system

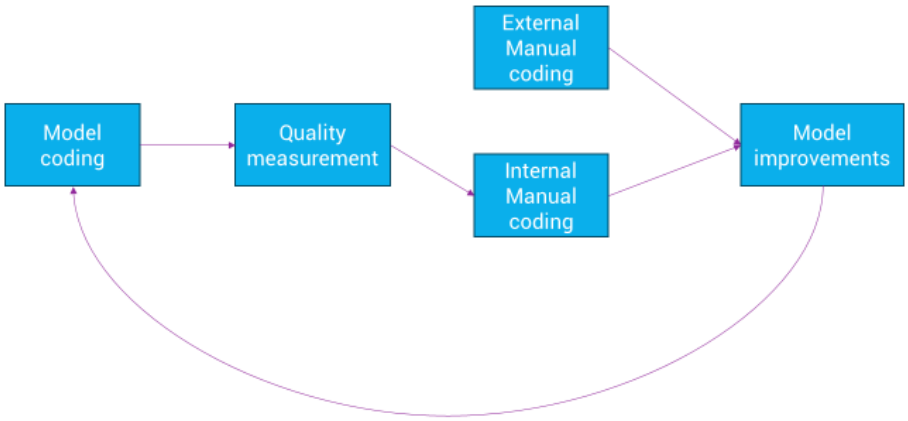
A broad use of the Swedish business register calls for an extensive quality assurance which may result in high costs or increased response burden. It may be necessary to prioritize. In the previous transition, subgroups of the population were pointed out as having different priority in the manual coding, but without an explicit level for the quality. These should be specified to carry out the quality assurance process effectively.

Additionally, quality levels need to be defined for specific codes or groups of codes (see Section 3.1.1). We suggest a minimum value of five for the absolute percentage error for the divisions in NACE i.e., a grouping of 2-digits NACE codes, and important sub codes. Remaining codes could have a minimum value of ten for the absolute percentage error.

3.4 Integration with the coding methods

Mainly, the idea of the quality assurance process is to integrate it to a high extent with the coding methods. This allows for continuous measurement and improvement of quality and facilitates an effective allocation of resources. Naturally, the first step is to automatically code every unit in the register with either the LLM, cluster analysis, or imputation model. The next suggested step is to use either Method 1 or Method 2 to estimate the quality measures. Together with the priority system, the quality estimates form the basis to decide which codes or units to allocate resources to and improve through Methods 3 or 4. New externally coded NACE Rev. 2.1 codes are continuously received from the online platform mentioned in Section 2.2.1. The process is illustrated in Figure 1.

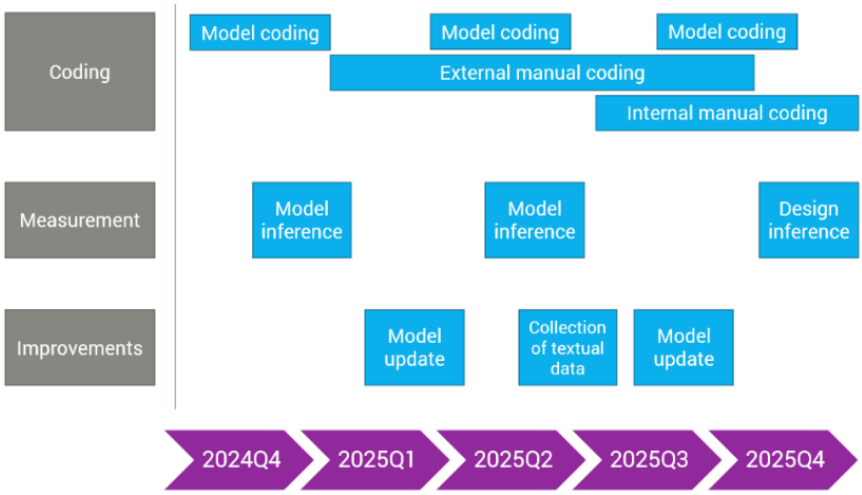
Figure 1: The overall quality assurance process.



3.5 Quality assurance workflow

In Figure 2, an example of the workflow of the quality assurance process is presented, conditional on the requirement that the NACE Rev. 2.1 codes need to be implemented in Sweden's business register by the end of 2025. The process facilitates an effective use of manual resources, which is shown through the late start of internal manual coding including design inference, and few updates of the model, which both may lead to manual labour. Furthermore, data collection could take place around the middle of the timeline. Thus, it is possible to only collect additional data for units with non-acceptable quality.

Figure 2. A potential workflow of the quality assurance process at Statistics Sweden.



4. Discussion

The presented process is built on existing continuous quality assurance processes. It focusses on coding with models and how to evaluate and improve the coding with an effective use of resources. Several methods for measuring quality are suggested to facilitate an effective process. A risk with this process is that cost of manual coding only transfers to other parts in the process, for example improvements of models. It is important to start with an idea for when to make improvements and how to carry them out automatically to a high extent. In this paper, a suitable workflow for Statistic Sweden is presented but other alternatives are possible.

Furthermore, the process requires several components i.e., data, quality indicators, coding methods, and a priority system. It is of essence that these components facilitate a result of good quality.

Lastly, it is of importance to underline the error of the current NACE Rev. 2.0 codes. Error occurred in the re-coding should be viewed in relation to the existing error in the business register to allocate the resources effectively. Overall, the goal should be to have a good quality for the entire business register and not only for the NACE codes affected by the revision.

References

Biemer, P. P., & Lyberg, L. E. (2003). Introduction to Survey Quality. John Wiley & Sons.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805. Retrieved from http://arxiv.org/abs/1810.04805

Malmsten, M., Börjeson, L., & Haffenden, C. (2020). Playing with Words at the National Library of Sweden – Making a Swedish BERT. arXiv:2007.01658. Retrieved from https://arxiv.org/abs/2007.01658

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. Artificial Intelligence Review, 56, 3005–3054. Retrieved from https://doi.org/10.1007/s10462-022-10246-w

Sande, L. S., & Zhang, L.-C. (2021). Design-Unbiased Statistical Learning in Survey Sampling. Sankhya. Series. A, 83(2), 714–744. https://doi.org/10.1007/s13171-020-00224-1

Wallgren, A., & Wallgren, B. (2022). Register-based Statistics: Registers and the National Statistical System. John Wiley & Sons.

UNECE Machine Learning Group. (2022). The quality of Training data. UNECE Statswiki. 1 Retrieved from Machine Learning Group 2022 - Machine Learning for Official Statistics - UNECE Statswiki

Eurostat. (2023). Handbook on implementation of NACE Rev. 2.1 in Business Registers.

Faria, T., & Seimandi, T. (2023). Classifying companies in France using machine learning. UNECE Machine Learning for Official Statistics Workshop 2023. Retrieved from https://unece.org/sites/default/files/2023-05/ML2023_S1_France_Faria_Paper.pdf

OpenAI. (2024). Guides - Vision. OpenAI Platform Documentation. Retrieved April 5, 2024, from https://platform.openai.com/docs/guides/vision